La Web de los Datos

Gentileza: Daniel Hernández.

Claudio Gutiérrez
Profesor Asociado DCC Universidad de Chile. Ph.D. Computer Science, Wesleyan University; Magíster en Lógica Matemática, Pontificia Universidad Católica de Chile; Licenciatura en Matemáticas, Universidad de Chile. Líneas de especialización: Fundamentos de la Computación, Lógica Aplicada a la Computación, Bases de Datos, Semántica de la Web.
cgutierr@dcc.uchile.cl

Daniel Hernández
Estudiante de Magíster en Ciencias de la Computación e Ingeniero Civil en Computación, Universidad de Chile. Entre sus áreas de interés se encuentran la Web, la publicación de datos y el acceso a la información pública.
daniel@degu.cl

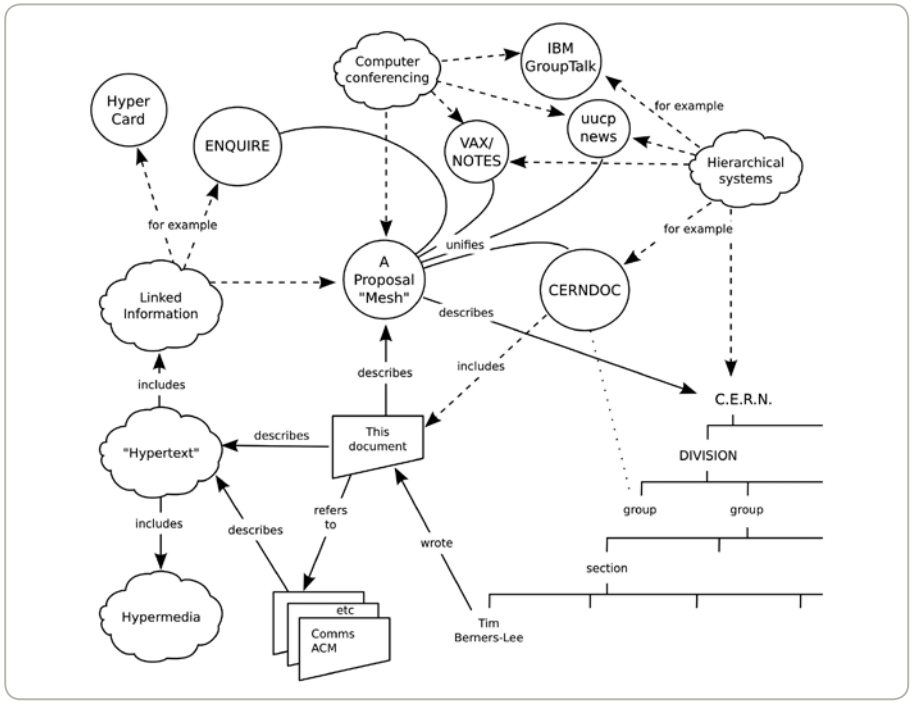
Desde el punto de vista de la información, probablemente la conceptualización más ingenua (pero también más entendible) de la Web sea la de una biblioteca infinita. La idea no es nueva y en 1939 ya Borges la había explicitado en su cuento La Biblioteca Total. *"Todo estaría en sus ciegos volúmenes"*, escribe. De hecho, concebir un espacio universal de información como una generalización de una biblioteca es muy útil. Incluye casi todas las facetas que uno querría que tuviera tal artefacto. Pero tiene un sesgo fundamental: la biblioteca está compuesta de libros, esto es, en términos de la Web, de documentos. Documentos son artefactos producidos por humanos para ser procesados ("consumidos") por humanos. Si uno reemplaza en este modelo el rol que juegan los libros (o documentos) por "datos", lo que se tiene es un modelo abstracto de lo que se llama la "Web de Datos". El cambio parece menor, pero sus consecuencias son impredecibles. Este es el objeto del cual nos ocuparemos en este artículo.

El diluvio de datos. Estudiar la Web de Datos es un tema muy relevante. Permítannos insistir sobre este punto. La gran expectación existente acerca de los inimaginables niveles de producción, disponibilidad y usos de datos (sensores, experimentos, ciencias, estadísticas, redes sociales, etc.) indican que estamos viviendo un cambio fundamental en las prácticas tradicionales de producción, intercambio y procesamiento de la información. La ola de datos fue observada ya hace algunos años por analistas de tecnologías. O'Reilly, en 2005, al tratar la Web 2.0 [1] indicaba que

"los datos son el siguiente Intel". En un nivel más académico, un informe de la comunidad internacional de Bases de Datos [2] advertía que la ubicuidad de "Grandes Datos" iba a remecer las bases de esta disciplina. Szalay and Gray, basados en que la cantidad de datos científicos se duplica cada año, hablaban en 2006 de "un mundo exponencial" [3] y Bell y sus colegas [4] lo llamaron "Diluvio de Datos". Todos se referían al fenómeno del incremento exponencial de volúmenes de datos comparado con el de una década atrás, debido a los avances tecnológicos que permiten capturarlos, transmitirlos y almacenarlos: satélites, telescopios, instrumentos de alto rendimiento, sensores, redes, aceleradores, supercomputadores, etc. Pero el fenómeno no es exclusivo de las áreas científicas. Tendencias similares pueden encontrarse en casi todas las áreas de la actividad humana. Las redes sociales están generando, no sólo grandes volúmenes de datos, sino también redes complejas que piden nuevas técnicas y enfoques para la gestión y procesamiento de datos. Las nuevas tecnologías también han impactado las políticas gubernamentales. Leyes de transparencia e iniciativas de publicación y archivo de datos están imponiendo el mismo tipo de desafíos al sector público [5]. Administrar, curar y archivar datos digitales se ha convertido en una disciplina *per se*. Algunos ya hablan de la "ciencia de los datos" [6]. Este fenómeno está impactando la disciplina de la computación en todas sus dimensiones, desde el nivel de sistemas, arquitecturas, comunicaciones, bases de datos, modelos de programación, ingeniería de software, etc. (ver por ejemplo: [7,8,9]). En todos estos desarrollos, la Web juega un rol central como una plataforma natural donde "viven" y se encuentran estos datos.

En este artículo exponemos sucintamente las iniciativas y tecnologías más relevantes que se han desarrollado para abordar los desafíos del manejo de datos en este nuevo escenario. Presentaremos primero las nociones básicas de la Web. Luego, abordaremos las dos iniciativas más relevantes en estos temas: Linked Data y Open Data. A continuación, presentaremos las técnicas actuales para la publicación y el acceso de datos abiertos. Finalmente, describimos algunas de las herramientas más importantes que se están usando en la Web de Datos.

Figura 1



La primera propuesta de la Web por Tim Berners-Lee. Nótese las ideas subyacentes: datos heterogéneos, usuarios heterogéneos, ausencia de jerarquías, redes, principalmente documentos (tomado de Tim Berners-Lee, Information Management: A Proposal).

BREVE DESARROLLO DE LA WEB

Tim Berners-Lee (TBL en adelante), el creador de la Web, la definió como "un espacio de información compartida a través del cual personas y máquinas se pudieran comunicar" [10]. En otra intervención, insistía que "lo más importante de la Web es que ella es universal" [11]. Veremos que esta universalidad está estrechamente ligada al compartir. No debe ser privativa de una compañía, ni de un gobierno, ni de una organización particular, sino que debe ser compartida por toda la gente alrededor del mundo.

El problema técnico que motivó el primer diseño de la Web, fue desarrollar un espacio para la gente que trabajaba en el CERN, que provenía de diferentes países, con diferentes costumbres, diferentes idiomas; manejando información muy heterogénea, como directorios de direcciones y teléfonos, notas de investigación, informes y mensajes, documentación oficial, etc., y basados en una infraestructura también heterogénea: terminales, servidores, supercomputadores, diversos sistemas operativos, software y formatos de archivos.

Roy Fielding [12], uno de los importantes arquitectos de los protocolos de la Web, resumía estos desafíos así: construir un sistema que debiera proveer una interfaz universalmente consistente a esta información estructurada, disponible en tantas plataformas como sea posible, y desplegada incrementalmente a medida que nueva gente y organizaciones se integren al proyecto.

Administrar, curar y archivar datos digitales se ha convertido en una disciplina *per se*. Algunos ya hablan de la "ciencia de los datos". Este fenómeno está impactando la disciplina de la computación en todas sus dimensiones.

En 2001 TBL [11] recordaba así los desafíos técnicos de tal proyecto:

El concepto de Web integraba diversos y distintos sistemas de información, por medio de un espacio imaginario abstracto en el cual las diferencias entre ellos no existan. La Web tenía que incluir toda la información de cualquier tipo sobre cualquier sistema. La única idea común que amarra todo era la noción de Identificador Universal de Recursos (URI), que identificaba un documento. A partir de allí, una serie de diseños de protocolos (como HTTP) y formatos de documentos (como HTML), que permitían a los computadores intercambiar información, traduciendo sus propios formatos locales en estándares que proveyeran interoperabilidad global.

Resumamos: la arquitectura de la Web se basa en tres pilares

1. **URI** (Universal Resource Identifiers): conjunto de identificadores globales que pueden ser creados y administrados en forma distribuida.
2. **HTTP** (Hyper Text Transfer Protocol): protocolo para intercambiar datos en la Web cuyas funcionalidades básicas son poner datos (put) y obtener datos (get) desde este espacio abstracto.
3. **HTML** (Hyper Text Markup Language): lenguaje para representar información y presentarla (visualmente) a humanos.

De estos tres, los identificadores globales son la base. TBL enfatiza esto diciendo que *"la Web fue diseñada para descansar sobre una especificación: los URI"*. La forma particular que tomaron el protocolo de transferencia (HTTP) y el lenguaje (HTML) fueron soluciones temporales con la tecnología disponible en ese tiempo.

PROTOCOLOS PARA LA WEB

La Web, tal como fue planteada por TBL, podría entenderse como un espacio donde se podría preguntar por URIs y recibir, como respuesta, documentos. Está de algún modo implícito que se espera recibir exactamente aquel documento que es identificado por la URI. No obstante, el protocolo es lo suficientemente abierto para poder implementar otras funcionalidades, por ejemplo, recibir documentos que dependan del usuario. Roy Fielding, de quien hablamos antes, es uno de quienes más ha avanzado en los requerimientos de protocolos Web, es decir, en la definición del comportamiento esperado para interoperar en ella. Por razones de espacio, mencionemos aquí sólo las restricciones que él sugiere en un modelo de arquitecturas que llama REST:

Cliente-servidor. Los clientes deben estar separados de los servidores por una interfaz uniforme. Esto permite modularizar el desarrollo y extensibilidad de las aplicaciones.

Ausencia de estado. Cada pedido de un cliente a un servidor debe contener toda la información necesaria para entender el requerimiento, y no debiera sacar provecho de ningún contexto almacenado en el servidor. El estado de la sesión debiera ser enteramente mantenido en el cliente.

Cacheable. Esto es, que los datos de una respuesta puedan ser implícitamente etiquetados como susceptibles de ser mantenidos o no el en caché. En caso de sí, se da el derecho a reusar esa respuesta para futuros pedidos equivalentes.

Interfaz uniforme. Una funcionalidad central que debiera distinguir la arquitectura de la Web de otras, es el énfasis en interfaces uniformes entre componentes.

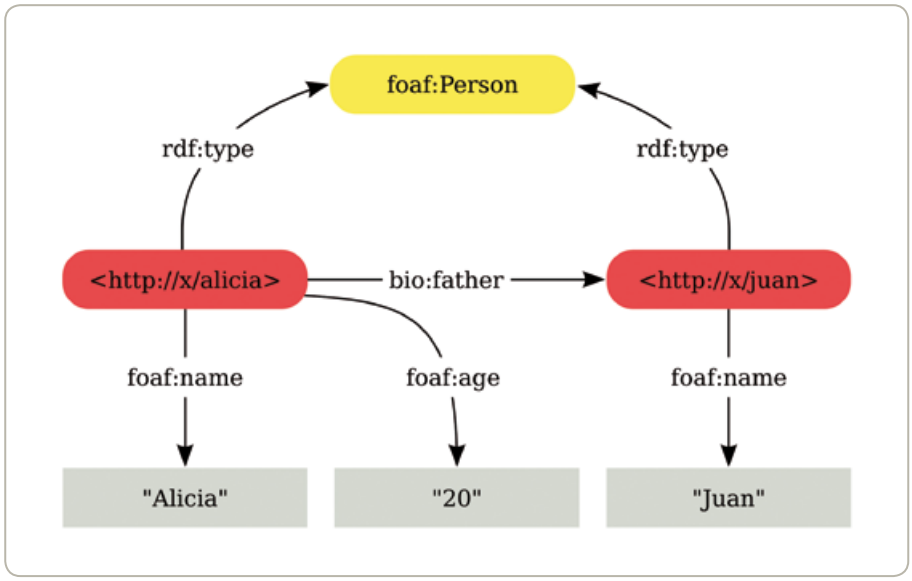
Sistema de niveles. Arquitectura compuesta de niveles jerárquicos, donde cada componente no puede "ver" más allá del nivel inmediato en que está operando.

LENGUAJES PARA LA WEB

Paradójicamente, una de las razones para el éxito de la Web fue la falta de semántica y estructuración del formato de sus documentos, el lenguaje HTML, que surgió más orientado a los elementos visuales que a la codificación de las estructuras de los documentos.

Una segunda generación, XML, permitió definir la estructura de los documentos con mayor precisión, representando los documentos como árboles y agregando reglas que permitieron establecer restricciones en la anidación de los elementos. Las distintas versiones del lenguaje de la Web han ido progresivamente separando la semántica de la presentación, al crear un lenguaje específico para definir la apariencia visual de los elementos, CSS, y retirar atributos que anteriormente permitían definir la apariencia (como @color, @width, etc.). Siguiendo esta tendencia, en las última versión, HTML5, se han incorporado elementos como aside, article, details, menu, nav, header, footer, etc. cuya función es netamente identificar semánticamente la estructura de los documentos.

Figura 2



Ejemplo de grafo RDF. El triple principal en rojo, representa la afirmación "Juan es padre de Alicia". Nótese que además podemos indicar propiedades de cada una de estas personas (rectángulos grises), por ejemplo, nombre y/o edad. Además, podemos incluir información del tipo de objeto de ambos recursos (en este caso son personas, esto es de tipo foaf:Person).

Sin embargo, a pesar del progresivo avance en la separación entre la información y apariencia, para la Web de los Datos esto no fue suficiente, pues el diseño del lenguaje aún tenía en mente el modelo de documento de texto diseñado para ser leído por humanos. Entonces, cabe preguntarse: ¿cuál es el "buen" lenguaje para la representación y el intercambio global de datos? He aquí algunos requerimientos básicos:

1. Que sea suficientemente flexible para describir la mayoría de los tipos de datos (en particular datos, metadatos y conocimiento).
2. Que sea minimalista y eficiente en lo referente a las necesidades de los usuarios y la complejidad de procesamiento.
3. Que pueda escalar en forma distribuida (no centralizada).

La Web Semántica. Hay dos desafíos que motivan una extensión natural de las ideas de la Web a un proyecto que se ha llamado la Web Semántica (en adelante WS): a) si los datos y la información escalan a niveles más allá de la capacidad normal de los humanos (como ocurre hoy día), la única posibilidad de accesarlos, organizarlos y administrarlos es vía automatización. b) El problema del significado de la información: ¿cuál es el significado de cada pieza de información en la Web? Esto tiene que ver fundamentalmente con la semántica y el significado de los conceptos (aún en el mismo lenguaje).

La WS intenta resolver estos problemas basada en la simple idea de organizar la información a nivel planetario. La WS es *"la Web de Datos procesables por máquinas"* escribe TBL. Y esto significa estandarizar significados. Para ello la WS utiliza un modelo de datos que se conoce como Resource Description Framework (RDF) [13] y que está basado en la forma básica de las oraciones, compuestas de sujeto, predicado y objeto. Estas tripletas *(s,p,o)* pueden ser entendidas como fórmulas lógicas binarias del tipo *p(s,o)*.

Un conjunto de tripletas puede ser interpretado como una red semántica, es decir, como un grafo dirigido con nodos y arcos rotulados, donde para cada triple hay un arco rotulado con el predicado y los nodos inicial y final son rotulados con el sujeto y el objeto. Así, la Figura 2 describe dos recursos que representan personas llamadas Alicia y Juan, donde Alicia es hija de Juan.

RDF no sólo describe una estructura de grafos, sino que en él también se definen los conceptos de clase e instancia. En el ejemplo de la Figura 2, los recursos que representan las personas son instancias de la clase foaf:Person.

El otro componente de la WS lo forman la capacidad de establecer reglas que permitan modelar (y validar modelos) y deducir afirmaciones (tripletas) a partir de otras. Para ello se definió The Web Ontology Language (OWL) [14], que es una codificación de la lógica en el lenguaje RDF, diseñado para describir ontologías y asociado a un semántica que define reglas de razonamiento para ellas.

En este punto podríamos detenernos brevemente para señalar una separación entre los caminos del desarrollo de RDF: el de representar datos mediante estructuras de grafos y el de introducir reglas de razonamiento. El primero se enfoca en la idea de una gran base de datos y, como consecuencia natural, requiere de lenguajes de consulta para ella, siendo el más popular SPARQL (una símil de SQL para datos RDF en la Web). El segundo, en cambio, visualiza la información como una base de conocimiento y por ende busca definir reglas para inferir conocimiento a partir de lo ya conocido.

Enmarcados en el compromiso usual entre la expresividad y la complejidad de procesamiento, se han desarrollado varios lenguajes para codificar vocabularios para RDF y por ende las reglas de inferencia que ellos otorgan a los datos expresados. Estos lenguajes pueden ser agrupados, en grueso modo, en tres grupos: a) aquellos con una mínima semántica o sin ella (esencialmente para definir jerarquías de tipos, clases y predicados) [15], b) RDF Schema más algunas extensiones menores y c) OWL, el lenguaje para las ontologías de la Web. No obstante, para enlazar y describir datos, a) pareciera ser suficiente.

LA WEB DE DATOS: LINKED DATA Y OPEN DATA

La Web de Datos es una colección global de datos producidos por la exposición y publicación sistemática y descentralizada de datos (crudos), usando protocolos y lenguajes de la Web.

Sobre la infraestructura de RDF

No es una sorpresa que la noción de la Web de los Datos esté estrechamente relacionada con la WS. Aquí brevemente presentaremos las fortalezas del modelo RDF y los desafíos que se enfrentan para abordar la Web de los Datos.

RDF fue diseñado para facilitar el procesamiento automático de la información en la Web por medio de metadatos. En 1999 la recomendación establecía con claridad: "RDF sirve para situaciones donde la información necesita ser procesada por aplicaciones, en vez de sólo ser desplegada para seres humanos". De este modo, el objetivo principal es la inclusión de información accesible por máquinas en la Web. Pero el diseño de RDF tiene otra consecuencia, su estructura de grafo permite la representación de una amplia gama de datos, abriendo la puerta a la conversión de la Web de los documentos a la Web de los Datos.

El poder de RDF nace de la combinación de dos ideas: a) un modelo flexible para representar tanto datos como sus metadatos de una manera uniforme, en la que ambos compartirían el mismo estatus de objetos de información. b) La estructura de grafos representa naturalmente las interconexiones y relaciones entre los datos. De hecho, esta última característica es la que sustenta el desarrollo de la iniciativa Linked Data.

Linked Data

Entre los proyectos más exitosos que atacan el problema de la ubicuidad de datos en la Web está Linked Data [16,17]. Los autores del proyecto lo definen así [16]:

Linked Data se trata de usar la Web para conectar datos relacionados que no han sido previamente enlazados, o usar la Web para disminuir las barreras para enlazar datos que hoy usan otros métodos. Específicamente, Wikipedia define Linked Data como "una buena práctica recomendada para exponer, compartir, y conectar piezas de datos, información, y conocimiento en la Web Semántica usando URIs y RDF".

La idea es simple: gracias a las tecnologías de la Web, es posible la producción, publicación y consumo de datos (no sólo de documentos) lo que se ha hecho universal. Sacar provecho de esto significa superar uno de los principales problemas hoy: el que estos datos están desconectados unos de otros, impidiendo su aprovechamiento conjunto.

TBL [18] explica como sigue las principales ventajas de Linked Data:

- **Permite conectar diferentes cosas de diferentes fuentes de datos.** El valor agregado de poner datos en la Web estriba en que se los puede consultar en combinación con otros tipos de datos de los cuales uno ni siquiera estaba consciente que existían.
- **Es descentralizado,** permitiendo que cada agencia y persona pueda crear y publicar sus propios datos, sin barreras editoriales, comerciales o administrativas.
- **Uso de estándares abiertos libre de licencias,** significa que nadie, agencias, gobiernos o personas, quedan ligados permanentemente a ningún proveedor.
- **Un círculo virtuoso.** Hay muchas organizaciones y compañías que se motivarán con la presencia de datos para desarrollar sobre ellos diversas aplicaciones y accesos a diferentes grupos de usuarios.

El mismo TBL propuso un test de "cinco estrellas" para la publicación de datos:

1. Ponga su material disponible en la Web (en cualquier formato).
2. Póngalo como datos estructurados (por ejemplo, Excel en vez de imagen escaneada de una tabla).
3. Use formatos no propietarios (por ejemplo, CSV en vez de Excel).
4. Use URLs para identificar cosas, de tal forma que la gente pueda apuntar (y referenciar) a su material.
5. Enlace sus datos con los de otra gente para proveer contexto.

Open Data

Datos abiertos (Open Data) es un movimiento que apunta a facilitar la producción y diseminación de datos e información a escala global. Debido a su relación con los temas que surgen de la discusión de lo "público versus lo privado", el movimiento ha llegado a ser muy influyente en la administración y manejo de la información en gobiernos, bibliotecas y grandes organizaciones.

Podemos definir Open Data de la siguiente manera: "Datos Abiertos es un movimiento cuyos objetivos es desarrollar y difundir estándares abiertos para los datos en la Web".

Por supuesto la gran pregunta es qué significa "datos abiertos". Seguiremos aquí el enfoque metodológico de Jon Hoem en su estudio de comunicación abierta [19], adaptándolo a nuestro ámbito. Hay muchas posibles dimensiones desde donde acercarse a la "apertura" de datos. Tres importantes son: el nivel de contenidos, el nivel lógico y el nivel físico. Para los datos, esto significa informalmente: semántica, tipos de datos y formatos, y hardware.

La gente ligada a datos gubernamentales es quien ha elaborado más a este respecto. Temprano, en 2007, se propusieron ocho principios para datos abiertos [20]. Aunque se refieren a "datos públicos", ellos ofrecen buenos puntos de vista genéricos:

1. **Que sean completos:** todos los datos deben estar disponibles.
2. **Que no estén procesados:** los datos se publican tal como fueron recolectados en la fuente, con el máximo nivel posible de granularidad (sin ser agregados ni modificados).
3. **Que sean actuales:** exponga los datos tan rápido como sea necesario para preservar su valor.

4. **Que sean accesibles:** hacerlos disponibles para el más amplio rango de usuarios y con los más diversos propósitos.
5. **Que sean susceptibles de automatización:** los datos razonablemente estructurados y marcados permiten su procesamiento automático o semiautomático.
6. **Que no haya discriminación:** los datos debes estar disponibles para todos sin necesidad de registrarse.
7. **Que no sean propietarios:** los datos deben estar disponibles en formatos para los cuales ninguna entidad tenga exclusivo control.
8. **Que sean licenciados abiertamente:** los datos no deben estar sujetos a ningún copyright, patente, marca registrada o regulación de secreto de negocio.

Los ocho principios anteriores definen lo que se puede considerar como Datos Abiertos, es decir, no implican que todos los datos deban cumplirlos. Muchas veces pueden existir buenas razones para no hacerlo, como la privacidad y la seguridad.

El impulso dado al desarrollo de modelos de Datos Abiertos ha descubierto varias actividades que eran consideradas como "dadas", o no habían ganado la prominencia que tienen hoy. Particularmente importantes aparecen actividades como preparar datos (para publicación), limpieza de datos, diseño de vocabularios internacionalizados, infraestructura física, disponibilidad de servicios, trazabilidad de origen, y particularmente, temas de licenciamientos y aspectos legales.

PUBLICANDO Y ACCESANDO DATOS ABIERTOS

Ambos proyectos, Linked Data y Open Data, son proyectos relativamente independientes. El primero busca entrelazar información generada y almacenada de manera distribuida y de naturaleza heterogénea con una tecnología apropiada, sitial que de momento es ocupado por RDF. El segundo hace hincapié en que los datos se encuentren disponibles para la ciudadanía, independientemente de la forma en la cual se puedan integrar.

En general las organizaciones se han visto enfrentadas ante la obligación de hacer pública su información. En Estados Unidos esta obligación surgió de una orden emanada desde la Presidencia y en el caso chileno comenzó con la Ley Nº 20.285, sobre el acceso a la información pública. Ante estas ordenanzas, los organismos que las tienen que cumplir se ven enfrentados ante los detalles técnicos y legales, sin poseer marco conceptual que les permita ejecutarlas adecuadamente. La falta de este marco para organizar, preservar y hacer que los datos públicos se mantengan accesibles en el largo plazo ha tenido como consecuencia que mucha información relevante desaparezca o que transcurra tiempo valioso con ella, fuera del alcance de quienes podrían haberla utilizado.

En los tiempos previos al advenimiento de las computadoras plantear metodologías para albergar la información resultaba una tarea sencilla de describir. R. A. Baker resumía las prácticas para mantener ordenados los apuntes de las investigaciones en laboratorios de química hacia 1933 de la siguiente manera [21]:

Dado que la investigación es un esfuerzo organizado para descubrir y una productiva aplicación de hechos, todos los datos obtenidos deben ser adecuadamente ordenados, correlacionados, interpretados y finalmente archivados con el fin de lograr el retorno del esfuerzo invertido. Cada experimento debiera ser titulado claramente y debería limitarse a una materia o variaciones de un sólo factor. El título debería aparecer al inicio de cada página dedicada al experimento. Luego del título inicial debería haber una descripción del problema, seguido del procedimiento, una descripción de los instrumentos, los datos y, finalmente, las conclusiones.

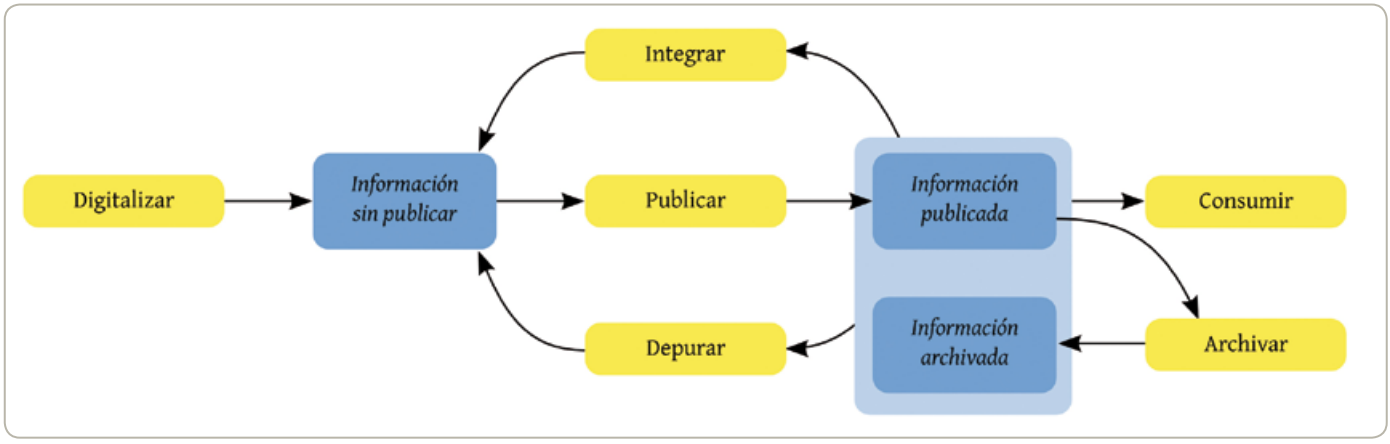
De igual manera las prácticas para mantener los archivos contables de una oficina, los acuerdos, las leyes, los archivos de bienes raíces, la información del registro civil, los registros médicos, etc. se definieron meticulosamente para satisfacer las necesidades de cada área y aprovechar las posibilidades que entregaba la documentación en papel. Sin embargo, con la llegada de las computadoras estas prácticas dejaron de tener vigencia. Las posibilidades de interrelacionar distintas fuentes de datos y de procesar de manera automática los crecientes volúmenes de información impusieron nuevos desafíos a la publicación de datos. En el caso particular de los datos científicos, las buenas prácticas que se ejemplifican en el flujo de datos que resumía R. A. Baker se dejaron de lado. Los medios actuales de comunicación de las investigaciones son los papers, pero estos rara vez van acompañados con referencias a los datos. Lo que es peor, en gran parte de los casos los datos no pueden ser accedidos, pues no siempre son públicos o porque se han desechado.

Sin duda los cambios introducidos por el uso de los computadores requieren hacer un cambio de paradigma en la manera en que tratábamos la información. Ello requiere detenernos a revisar y conceptualizar el proceso de la generación, preservación y uso de los datos en nuestros tiempos digitales.

Datos, datasets, archivos, bases de datos y distribuciones

La primera pregunta que salta frente a nosotros es qué son los datos. En general pareciera haber cierta convención tácita de que los datos deben ser los átomos de la información. Así, pareciera un buen acuerdo concluir que los datos son afirmaciones instanciadas, es decir, expresiones de la forma $a(x,y,...,z)$, donde a es una afirmación sobre los objetos $x,y,...,z$. Esta noción de datos es útil porque podemos llevarla fácilmente a nuestros espacios conocidos de las bases de datos relacionales y al modelo de triples de RDF. En el modelo relacional, cada fila de una tabla t puede ser entendida como una fórmula $t(x,y,...,z)$, donde los parámetros son los valores de las columnas en dicha fila. De manera similar en el modelo RDF, cada triple (s,p,o) puede ser entendido como una fórmula $p(s,o)$.

Figura 3



Acciones y estados de la información.

Con este concepto de datos nos resulta sencillo definir un dataset como un conjunto de datos que puede ser definido por extensión, enumerando todos los datos, o por comprensión, cuando podemos acotarlo de alguna manera aunque luego no podamos enumerar todos los datos. Por ejemplo, el conjunto de datos de todos los nacimientos en Chile durante 2010, es un conjunto que podríamos poner por extensión, mientras que el conjunto de las edades de todos los chilenos no, pues es algo que va cambiando y cualquier enumeración quedará rápidamente obsoleta. En lo siguiente, a los datasets expresados por extensión los llamaremos datos muertos, mientras que a los otros, datos vivos.

La clasificación entre datos vivos y muertos, nos facilita la diferenciación entre bases de datos y archivos. Un archivo es una secuencia de bits que podemos guardar o enviar por la red. En particular un archivo puede codificar un conjunto de datos por extensión, pero no uno por comprensión. Por otro lado, muchos datasets definidos por comprensión corresponden a bases de datos, cuyos contenidos cambian constantemente. Para interrelacionar ambos conceptos, podemos observar que el *dump* de una base de datos es siempre un archivo.

Por último, un concepto introducido por ontologías para catálogos de datos como DCat es el de distribución. Una distribución de datos es un medio por el cual podemos acceder a un dataset. En la Web las distribuciones son identificadas por URIs, que nos permiten descargar el archivo correspondiente a un dataset, cuando éste es expresable por extensión, o acceder a una interfaz que nos permite consultar la base de datos que lo define.

Actores y procesos en la vida de los datos

En una primera aproximación al mundo de los datos podemos suponer dos actores: quienes publican la información y quienes la usarán. No obstante, los roles que encontramos en los participantes son más variados y muchas veces los agentes participan cumpliendo más de un rol. La Figura 3 grafica un modelo algo más detallado de los roles de los actores definidos por sus actos (amarillos) y los estados por los que la información pasa (azules) como resultado de dichos actos.

Actualmente tenemos una pérdida entre la digitalización y el consumo. No todos los datos generados se encuentran disponibles para su consumo. Las preguntas son: ¿dónde se están perdiendo los datos? ¿Por qué y cómo podemos evitar que esto suceda? Refiriéndose específicamente a lo que ocurre con los datos científicos, Michael Witt le llamó a esta pérdida "information bottleneck" [22]. Como se mencionó inicialmente, el aumento en la capacidad de digitalización nos llevó al fenómeno referido como el diluvio de datos. El cuello de botella de información se encuentra entre la información sin publicar y la que está disponible para el consumo, es decir, en el proceso de publicación (ver Figura 3). Este proceso de publicación, también conocido como curación de datos, va desde definir estructuras y modelos apropiados para la información hasta generar identificadores para la información publicada y asegurarse de que ella quede accesible para el consumo. Las tareas de depuración e integración, dibujadas como procesos independientes en la Figura 3, pueden también encontrarse en el proceso de publicación en la medida que se busca agregar valor a los datos a publicar.

Dado el gran volumen de la información disponible para ser publicada, el consumo también presenta un desafío que debe ser facilitado en la publicación de los datos. De este modo, la publicación debe facilitar la automatización de procesos tales como: encontrar fuentes de información, buscar información dentro de ellas, extraer partes, integrar y visualizar.

Integración

Es uno de los mayores desafíos que impone la publicación de datos en la Web. La integración de datos consiste en proveer a los usuarios (o consumidores) una interfaz común para acceder transparentemente a datos dispersos y de naturaleza heterogénea [23]. Por ejemplo, un hipotético servicio que

recoge datos de pronósticos meteorológicos provenientes de la Dirección Meteorológica de Chile (meteochile.cl), los contrasta con un servicio extranjero como The Weather Channel (weather.com) y, además, entrega información sobre la disponibilidad hotelera en las distintas localidades.

Además la integración, es el núcleo de los problemas que se busca resolver con la iniciativa Linked Data. En RDF se proponen las URIs como elemento para identificar recursos y cumplen un rol fundamental en la manera que es posible referirse a recursos comunes desde datasets distintos.

Sin embargo, utilizar URIs no basta, la integración requiere que éstas sean compartidas entre los diferentes datasets. Esto involucra también a aquellas URIs que forman parte de los vocabularios, es decir, aquellas que identifican predicados, clases e instancias de uso común (ejemplo: bio:father, foaf:Person, dbp:Chile).

En vez de definir vocabularios propios, comúnmente se recomienda reutilizar vocabularios existentes con el fin de favorecer la interoperabilidad de la información publicada. No obstante, en algunos casos no resulta posible encontrar vocabularios existentes que se adapten a los datos, ya sea por la inexistencia de vocabularios para describir un área demasiado específica o porque nuestros datos poseen localismos que difieren de los modelos conceptuales que, en su mayoría, son diseñados para culturas que difieren de la nuestra. Aún en estos casos suele ser preferible extender vocabularios existentes a crear vocabularios desde cero.

Las recomendaciones anteriores, se deben en gran medida a que aún no está resuelto el problema de cómo integrar datos expresados con distintos vocabularios. Algunas estrategias para enfrentar este problema son: a) traducir los datos de un vocabulario a otro antes de consultarlos, b) aplicar reglas de deducción al momento de realizar la consulta y c) modificar la consulta de modo que permita trabajar con datos expresados en más de un vocabulario. No obstante, el problema de la integración de datos que usan distintos vocabularios en la Web es un problema abierto.

> Sin duda los cambios introducidos por el uso de los computadores requieren hacer un cambio de paradigma en la manera en que tratábamos la información. Ello requiere detenernos a revisar y conceptualizar el proceso de la generación, preservación y uso de los datos en nuestros tiempos digitales.

Otra barrera a la integración de los datos es la ausencia de un modelo universal de la información. Antero Taivalsaari [24] lo resume brevemente:

Un ejemplo de un concepto que es difícil de definir en términos de propiedades compartidas es "obra de arte". Ya que nadie puede definir límites claros para qué es arte y qué no lo es, no hay ninguna clase general "obra de arte", que comparta propiedades comunes. La definición es subjetiva y depende en gran medida de la situación o del punto de vista.

Algunas personas viviendo cerca del Ecuador no pueden distinguir entre hielo y nieve, mientras los esquimales tienen numerosas palabras para distinguir entre distintos tipos de nieve. Los Dani, de Nueva Guinea, tienen sólo dos términos de colores básicos: mili (oscuro / frío) y mola (luminoso / cálido) que cubre el espectro completo, y tienen gran dificultad para diferenciar entre colores con mayor detalle.

Los lenguajes para definir vocabularios RDF Schema y OWL se fundamentan en la definición de clases y subclases, lo que implica establecer jerarquías entre ellas. Las observaciones de Taivalsaari ponen en duda que tal construcción pueda extenderse a nivel planetario. Un fenómeno similar puede visualizarse en bibliotecología, donde las grandes jerarquías ceden paso a pequeños tesauros funcionales que pueden aplicarse simultáneamente para describir un mismo conjunto de recursos. Siguiendo la estrategia de los pequeños tesauros, Simple Knowledge Organization System (SKOS) es un lenguaje que permite definir esquemas conceptuales para ser aplicados independientemente unos de otros, sin requerir la construcción de una jerarquía única.

HERRAMIENTAS PARA PUBLICAR

Diversas herramientas han surgido a la par con las necesidades identificadas en la práctica de la publicación de datos. La mayoría de los organismos públicos que tomaron el desafío de hacer accesible la información pública a la ciudadanía comenzaron con catálogos de datos, donde los datasets, al igual que los catálogos de documentos, eran tratados como objetos opacos, en los cuales sólo es posible acceder de manera uniforme a ciertos metadatos comunes. Los catálogos cumplen con el objetivo básico de hacer accesibles y referenciables a los datasets, pero aún presentan una deuda: la integración de datos. Es allí donde el modelo RDF entra en juego, proveyendo de herramientas para integrar lógicamente los datos y para consultarlos. Aunque aún quedan temas abiertos, como el balance entre la centralización y la distribución.

Catálogos

La creciente publicación de datos por gobiernos y organismos públicos se ha realizado mayoritariamente en la forma de catálogos de datos. La publicación de catálogos nacionales de datos fue impulsada con el precedente establecido por los Gobiernos de Estados Unidos y Reino Unido, con sus catálogos lanzados en mayo de 2009 y en enero de 2010, respectivamente. En un corto período de dos años ya han surgido numerosos catálogos de gobiernos locales, regionales y nacionales, así como también de organizaciones internacionales como el Banco Mundial y numerosas ONG. Existen varias organizaciones preocupadas de hacer una suerte de metacatálogo, es decir, listar y describir todos los catálogos de datos existentes, entre ellos destacan los de la fundación CTIC, la Open Knowledge Foundation (OKF) y el Resselaer Polytechnic Institute (RPI). En el más reciente recuento, la OKF contabiliza la existencia de 139 catálogos de datos.

En Chile, si bien existen varios organismos públicos que están dejando disponible la información, aún no se ha logrado lanzar un portal que permita un acceso común a todas las fuentes de datos nacionales, por lo que muchas de ellas son desconocidas para la población. Junto con las dificultades de encontrar, la información publicada por la mayoría de los organismos públicos chilenos suele encontrarse en formatos que dificultan su procesamiento automatizado e integración con otras fuentes de datos.

Un catálogo puede entenderse como una colección de entradas describiendo conjuntos de datos, también conocidos como datasets. La descripción de los datasets suele incluir metadatos tales como el nombre, la descripción, las materias tratadas, el origen, la fecha de publicación, las licencias de uso, etc. Entre estos metadatos resultan fundamentales las referencias para poder acceder a los datos. En algunos casos estas referencias son teléfonos o direcciones para consultar por ellos, como en el caso del catálogo de datos geográficos de Chile, mantenido por el Servicio Nacional de Información Territorial (SNIT). No obstante, cuando se habla de catálogos de Datos Abiertos lo esperable es que éstos sean accesibles a través de la Web, ya sea mediante documentos descargables (datos muertos) o servicios que permiten consultar por datos en línea (datos vivos).

La gran aceptación que ha ganado el proyecto Linked Data, ha influido en que algunos catálogos modelen y publiquen la información de los datasets con RDF. Un ejemplo de ello es el catálogo de Australia, donde las páginas del catálogo se encuentran en formato RDFa, una extensión de XHTML que permite marcar datos usando el modelo RDF. Para el catálogo de datos públicos del Gobierno australiano se creó un vocabulario RDF específico para expresar sus metadatos, el AGLS, aunque también existen vocabularios de uso general para catálogos como DCat y VoID. El primero de ellos es aplicable para catálogos donde los datos pueden ser publicados en cualquier medio, mientras que el segundo, es específico para interrelacionar datasets publicados en el modelo RDF, ya sea a través de archivos o servicios de consulta (SPARQL endpoints).

El uso de catálogos para dataset responde principalmente a la necesidad de encontrar fuentes, mencionada al inicio de la sección "Actores y procesos en la vida de los datos" de este artículo, y entregarles identificadores que permitan agregar metadatos a los dataset. Así por ejemplo, el problema de los identificadores de datasets publicados de manera distribuida es resuelto por el proyecto Dataverse, utilizando el Universal Numeric Fingerprint (UNF), un identificador generado aplicando una función sobre el dataset con una muy baja probabilidad de colisionar. No obstante lo anterior, el problema de integrar los datos no es abordado en los catálogos, pues ellos se sitúan en un nivel en el cual los datos son visualizados como objetos oscuros de los que sólo se puede agregar información por medio de metadata.

Como comentamos anteriormente, una de las cualidades que hacen relevante a RDF es su modelo flexible para representar tanto datos como sus metadatos de una manera uniforme, en la que ambos comparten el mismo estatus de objetos de información. Así pues, la siguiente herramienta que describiremos, los SPARQL stores, se enfrentará directamente con el problema de la integración.

SPARQL stores

En general hablamos de RDF stores o triplestores para referirnos a bases de datos orientadas a almacenar y consultar datos en forma de triples RDF. En particular, hablamos de SPARQL stores cuando el lenguaje de consulta es SPARQL.

SPARQL es un lenguaje de consulta basado en patrones, es decir, para obtener un conjunto de recursos que satisfacen ciertas propiedades debemos establecer un patrón por medio del cual estos recursos se encontrarán en el espacio de información sobre el cual queremos buscar. Como el espacio de RDF corresponde a grafos, los patrones serán definidos con grafos. Por ejemplo:

```
SELECT ?a, ?b
FROM <http://x/grafo>
WHERE {
 ?a rdf:type foaf:Person .
 ?b rdf:type foaf:Person .
 ?c rdf:type foaf:Person .
 ?a bio:father ?c ;
 ?c bio:father ?b ;
}
```

Busca a todos los pares de nodos *(?a,?b)* dónde *?b* es el abuelo paterno de *?a*. Los elementos *?a, ?b y ?c* son las variables dentro del patrón que corresponde a lo que se encuentra entre los paréntesis que acompaña al WHERE. Las variables deben ser instanciadas para entregar la respuesta que se pide en el SELECT. Por último, FROM especifica el grafo desde donde deben tomarse los triples que se usarán para hacer calzar el patrón.

La noción de grafo, identificable mediante URIs, permite agregar metadatos a estos grafos, visualizándolos como datasets. Esto es especialmente relevante para manejar la proveniencia (linaje) de los datos, porque en muchos casos éstos podrían provenir de diversas fuentes, con distintas calidades, temática y usos de vocabularios.

A pesar de que los SPARQL endpoints satisfacen la necesidad de consultar e integrar datos, actualmente no eliminan la tensión entre concentrar información localmente para consultarla y publicar distribuidamente. Jeni Tennison explica esta situación [25]:

Lo que no me queda muy claro es cómo esta publicación distribuida de datos puede conciliarse con el uso de SPARQL para consultar. Después de todo, SPARQL no soporta (en la actualidad) la capacidad de realizar búsquedas federadas. De este modo, el uso de SPARQL sobre todos los datos enlazados distribuidos suena como si necesitáramos un triplestore central que contenga todo lo que querríamos consultar.

Publicación de RDF en la Web

Cuando hablamos de publicación de RDF en la Web podemos diferenciar principalmente entre dos estrategias: proveer un servicio de consulta para acceder a los datos directamente, como resulta mediante un RDF store y publicar los datos tal cual, en archivos que sea posible descargar.

La publicación de datos como archivos en la Web tiene dos variantes: a) publicarlos como archivos aparte de los diseñados para la visualización humana y b) usar las mismas páginas Web como soporte para la publicación de datos.

En la primera variante nos encontramos con las diversas serializaciones (sintaxis) de RDF como RDF/XML, N3, Turtle, TriG, TriX, etc. En la segunda variante consideramos lenguajes de marcado como RDFa, eRDF y Microdata, que resultan interesantes, pues permiten publicar datos haciendo pequeñas modificaciones en las plantillas, que generan las visualizaciones de los contenidos.

Por último, los desafíos de almacenar e intercambiar grandes volúmenes de datos llevan a plantear formatos de archivos compactos y que sean capaces de resumir de manera autocontenida lo que contienen, como es el caso de HTC [25]. BITS

REFERENCIAS

[1] T. O'Reilly, What Is Web 2.0, 2005.http://oreilly.com/web2/archive/what-is-web-20.html.

[2] R. Agrawal et al., The Claremont Report on Database Research, 2008. http://db.cs.berkeley.edu/claremont/.

[3] A. Szalay, J. Gray, Science in an Exponential World, Nature, Vol. 440, marzo de 2006, pp. 413–414.

[4] G. Bell, T. Hey, A. Szalay, Beyond the Data Deluge, Science, Vol. 323, marzo de 2009, pp. 1297–1298.

[5] DATA.gov, proyecto de publicación de datos del gobierno de Estados Unidos. http://www.data.gov/

[6] Mike Loukides, What is data science?, 2010. http://radar.oreilly.com/2010/06/what-is-data-science.html

[7] G. Bell, J. Gray, A. Szalay, Petascale Computational Systems: Balanced CyberInfrastructure in a Data-Centric World, Computer, Vol. 39, Issue 1, enero de 2006, pp. 110–112.

[8] M. Stonebraker, S. Madden, D. J. Abadi, S. Harizopoulos, N. Hachem, and P. Helland, The end of an architectural era: (it's time for a complete rewrite), Proc. VLDB '07, 2007. pp. 1150–1160.

[9] No SQL, http://nosql-database.org/

[10] T. Berners-Lee. WWW: Past, present, and future. IEEE Computer, 29(10), octubre de 1996, pp. 69–77.

[11] T. Berners-Lee. Commemorative Lecture The World Wide Web - Past Present and Future. Exploring Universality. Japan Prize Commemorative Lecture, 2002 http://www.w3.org/2002/04/Japan/Lecture.html

[12] R. T. Fielding, Architectural Styles and the Design of Network-based Software Architectures. Doctoral dissertation, University of California, Irvine, 2000. http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[13] G. Klyne, J. Carroll, Resource Description Framework (RDF) Concepts and Abstract Syntax, W3C Recommendation, 2004. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[14] D.L. McGuinness, F. van Harmelen, OWL Web Ontology Language Overview, W3C Recommendation, 10 de febrero de 2004, http://www.w3.org/TR/owl-features/

[15] S. Muñoz, J. Pérez, C. Gutiérrez, Simple and Efficient Minimal RDFS. J. Web Sem. 7(3), 2009.

[16] LinkedData Project, http://www.linkeddata.org

[17] Ch. Bizer, T. Heath, T. Berners-Lee, Linked Data - The Story So Far, International Journal on Semantic Web and Information Systems, Vol. 3, 2009, pp. 1-22.

[18] T. Berners-Lee, Linked Open Data. What is the idea?, http://www.thenationaldialogue.org/ideas/linked-open-data

[19] J. Hoem, Openness in Communicaton, First Monday, Volume 11, Number 7, 3 de julio de 2006.http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1367/1286

[20] Seminar on Open Government Data (Open Government Working Group), 7 y 8 de diciembre de 2007. http://resource.org/8_principles.html

[21] Baker, R. A. In the research laboratory. Journal of Chemical Education, Vol. 10, 1933, pp. 408–411.

[22] M. Witt, Institutional Repositories and Research Data Curation in a Distributed Environment, Library Trends, 57(2), 2009. http://docs.lib.purdue.edu/lib_research/104/

[23] T. Lee, Attribution Principles for Data Integration: Policy Perspectives, febrero de 2002.

[24] A. Taivalsaari, Classes vs. Prototypes - Some Philosophical and Historical Observations, Journal of Object-Oriented Programming, 1996.

[25] Jeni Tennison, Distributed Publication and Querying, blog personal. http://www.jenitennison.com/blog/node/143

[26] J. Fernández, C. Gutiérrez, M. Martínez-Prieto, Compact Representation of Large RDF Data Sets for Publishing and Exchange, ISWC 2010. LNCS 6496, pp. 193–208. Shanghai, China, 7–11 November 2010.